Initiative face au virus. Regards croisés sur l'épidemie de Covid-19 apportés par les données sanitaires et de géolocalisation (mars à octobre 2020)

Jamal Atif, Bertrand Cabot, Olivier Cappé, Olga Mula, Rafael Pinot

▶ **To cite this version:**

Jamal Atif, Bertrand Cabot, Olivier Cappé, Olga Mula, Rafael Pinot. Initiative face au virus. Regards croisés sur l'épidemie de Covid-19 apportés par les données sanitaires et de géolocalisation (mars à octobre 2020). [Rapport de recherche] Université PSL; Inria; CNRS. 2020. hal-03084832

HAL Id: hal-03084832
https://hal.science/hal-03084832

Submitted on 21 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

20 décembre 2020

INITIATIVE FACE AU VIRUS

RAPPORT #2

PÉRIODE MARS-OCTOBRE 2020

REGARDS CROISÉS SUR L'ÉPIDÉMIE DE COVID-19 APPORTÉS PAR LES DONNÉES SANITAIRES ET DE GÉOLOCALISATION

1. INTRODUCTION

Faisant suite au premier rapport sur l'analyse de la mobilité pendant la première vague de l'épidémie de Covid-19, ce second rapport a pour but d'apporter des éclairages sur la dynamique épidémique de la seconde vague. Deux questions principales sont abordées. La première concerne l'impact des mesures gouvernementales (confinement, déconfinement, couvre-feu, etc.) sur la dynamique de propagation du virus. Pour ce faire, nous avons conduit une analyse rétrospective des différentes phases de l'épidémie de mars à octobre 2020 à partir d'une modélisation statistique robuste permettant de détecter les changements de dynamique dans les données d'hospitalisations, aux échelles nationale, régionale ou départementale. La deuxième question porte sur le rôle de la densité de population dans la dynamique de la propagation du virus. Pour répondre à cette question, nous avons, dans un premier temps, construit des indicateurs permettant de mesurer les évolutions de la distribution des zones densément peuplées sur le territoire français en exploitant les données de géolocalisation fournies par Facebook dans le cadre du programme "*Data for Good*", puis mis en relation ces indicateurs avec les résultats de la modélisation statistique des différentes phases de l'épidémie.

De cette étude résultent les enseignements suivants. Pendant la première vague de l'épidémie, les dates de changement de la dynamique épidémique peuvent être directement mises en relation avec les mesures gouvernementales, le changement le plus marqué étant celui lié à la mise en place du premier confinement. On observe un tournant à partir de la seconde moitié du mois d'août.

L'analyse aux échelles régionale et surtout départementale des changements intervenus pendant l'été et au mois de septembre met en évidence de fortes inhomogénéités territoriales, avec une dynamique épidémique portée par des territoires souvent peu impactés lors de la première vague. L'analyse des données de géolocalisation confirme des déplacements importants (avec un pic atteint mi-août) de la population des régions du Nord de la France et des grandes métropoles vers les régions côtières. Pendant la période estivale, une partie importante du territoire a vu sa densité de population augmenter de plus de 25% par rapport à la date de référence prise au 22 mars (pendant le confinement) avec des zones de forte densité habituelles (grandes agglomérations) qui se sont dépeuplées au profit de certaines zones littorales.

S'agissant de l'impact de la densité de population sur la dynamique épidémique, les conclusions sont plus mitigées. Si la présence de zones de forte densité sur un territoire apparaît bien comme facteur corrélé à la vitesse de propagation de l'épidémie durant la première vague, la situation est plus complexe pour la deuxième vague de l'épidémie. On observe néanmoins sur l'ensemble de la période que les plus forts taux d'hospitalisations rapportés au nombre d'habitants sont fréquemment observés dans les territoires les plus densément peuplés.

A PROPOS

Face Au Virus est une initiative scientifique ouverte lancée par des chercheurs et chercheuses de l'Université PSL, du CNRS et d'Inria, dans l'objectif de contribuer à l'information générale sur l'épidémie ainsi qu'à la définition et au suivi des politiques publiques de gestion de crise.

Face Au Virus travaille sur la question centrale de la prise en compte des mobilités dans la modélisation de l'évolution de l'épidémie. Pour compléter les données statistiques sur l'épidémie collectées par les pouvoirs publics, l'initiative dispose de données, fournies par des acteurs privés (dont Facebook), qui permettent de suivre l'évolution de différentes caractéristiques des mobilités.

Face Au Virus porte une grande attention à ce que l'utilisation des données se fasse dans le respect des droits et des libertés fondamentales et ne vise en aucun cas à l'identification de comportements individuels. L'initiative ne collecte aucune donnée spécifique et ne traite que des résumés statistiques des données de mobilités.

2. ÉVOLUTION DE L'ÉPIDÉMIE DE MARS À OCTOBRE 2020

2.1 A L'ÉCHELLE NATIONALE

Pour décrire et analyser l'historique de l'épidémie de Covid-19 en France, nous avons choisi dans ce document de considérer principalement les données d'admission hospitalières qui sont consolidées depuis le début du mois de mars 2020. Sous l'hypothèse que la proportion d'hospitalisations parmi les personnes infectées par le virus de la Covid-19 est resté stable depuis le début de l'épidémie, ces données fournissent une information précieuse, et de qualité comparable dans le temps et sur l'ensemble du territoire national, concernant l'évolution de la population touchée par le virus. Néanmoins, les nombres d'hospitalisations étant relativement faibles, surtout à l'échelle des départements, l'extraction d'informations fiables et interprétables à partir de ces données nécessite l'utilisation de méthodes statistiques. Le modèle utilisé ici repose sur l'identification de périodes pendant lesquelles les nombres d'admissions hospitalières suivent une tendance de croissance (ou décroissance) exponentielle. L'identification de ces périodes ainsi que des taux de croissance correspondants se fait de façon robuste en utilisant une modélisation statistique des données qui prend en compte, notamment, les effets hebdomadaires récurrents (voir l'annexe 2 pour les détails des méthodes utilisées).

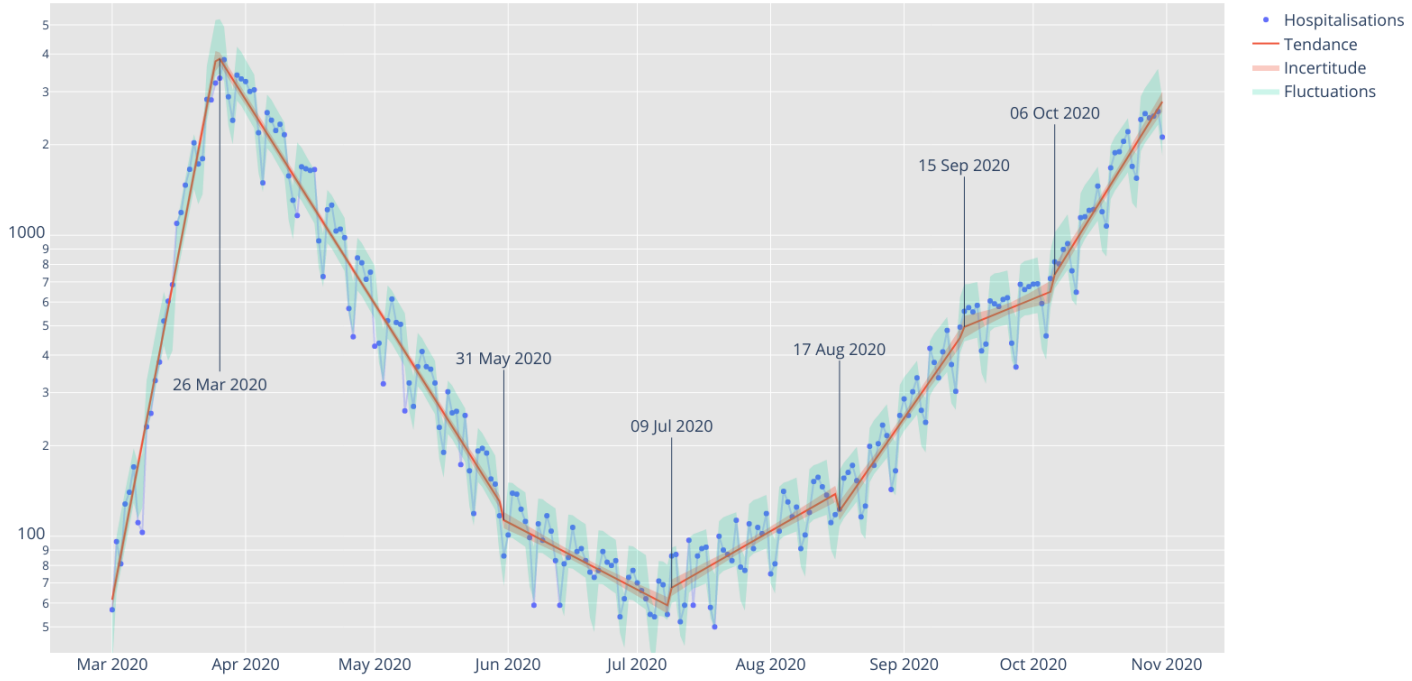


Figure 2.1 Modélisation statistique des données d'hospitalisation à l'échelle nationale

La figure 2.1 représente la modélisation la plus vraisemblable des données agrégées au niveau national. Les données sont représentées en échelle logarithmique, dans laquelle les périodes de croissance/décroissance exponentielle apparaissent comme des segments linéaires. Cette échelle logarithmique permet également de voir des effets qui seraient difficilement perceptibles autrement dans les périodes de ralentissement de l'épidémie (au creux de l'épidémie, autour du 9 juillet, le nombre d'hospitalisation est près de 100 fois plus faible que lors du pic de l'épidémie en mars). La figure représente également l'incertitude sur les tendances estimées (en rouge) ainsi que les fluctuations expliquées sur les données d'hospitalisation (en vert).

La figure 2.1 montre une bonne adéquation entre les données observées et le comportement modélisé. Les périodes successives sont caractérisées par des taux de croissance qui indiquent la vitesse de propagation de l'épidémie ainsi que par dates qui correspondent à des changements de comportement. Pour interpréter les événements qui peuvent être à l'origine de ces changements il faut prendre en compte le fait qu'ils se produisent avec un retard d'une quinzaine de jours. **Le changement le plus marqué, identifié en date du 26 mars, peut être rattaché à la mise en œuvre des premières mesures de confinement à partir du 12 mars**. De même, l'inflexion de la fin mai est consécutive au début de la levée du confinement, à partir du 11 mai, et la date du 9 juillet peut être reliée à la levée de mesures restrictives, concernant en particulier la région parisienne, le 22 juin (voir Table 2.2).

A partir du 9 juillet, on observe une remontée du nombre d'hospitalisations avec plusieurs régimes successifs qui, cette fois, ne semblent pas pouvoir être mis directement en relation avec les mesures gouvernementales mises en place dans la quinzaine précédente, en particulier en ce qui concerne la nette accélération constatée à partir de la mi-août. La date du 15 septembre, qui correspond à une courte période de relatif ralentissement de l'épidémie, peut être reliée à celle du 27 août où des mesures de contrôle ont été annoncées (notamment la généralisation du port du masque obligatoire dans l'ensemble des espaces fermés). Néanmoins, l'examen des données à des échelles plus fines (régionale et départementale) ainsi que le croisement avec des données de géo-localisation permet de formuler une autre hypothèse liée aux mouvements significatifs de population qui ont eu lieu pendant l'été.

Table 2.1 Facteurs de croissance estimés et leur traduction en nombres de de reproduction effectifs

Périodes	01 Mars - 26 Mars	26 Mars - 31 Mai	31 Mai - 9 Juillet	9 Juillet - 17 Août	17 Août - 15 Sept.	5 Sept. - 6 Oct.	6 Oct. - 31 Oct.
Facteur de croissance à 14 jours	10 - 12	0.4 - 0.5	0.7 - 0.9	1.2 - 1.4	1.8 - 2.1	1.1 - 1.4	1.9 - 2.3
Nombre de reproduction (effectif)	2.0 - 2.2	0.6 - 0.7	0.8 - 0.9	1.1 - 1.2	1.2 - 1.4	1. - 1.2	1.3 - 1.4

En ce qui concerne la vitesse de propagation de l'épidémie, les paramètres estimés peuvent être présentés sous la forme de facteurs de croissance à 14 jours (un facteur supérieur à 1 impliquant une augmentation du nombre de cas). Ce facteur semble particulièrement pertinent dans la mesure où un délai de 14 jours est nécessaire pour constater les changements produits par les mesures de contrôle de l'épidémie. La table ci-dessus présente les valeurs estimées avec leur incertitudes ainsi que leur traduction en terme de nombre de reproduction (en supposant une durée moyenne de la période de contagion du coronavirus égale à 6 jours). On observe que depuis la mi-août l'épidémie s'est remise à croître à un rythme qui, bien que très inférieur à celui observé dans la phase initiale, est assez significatif (doublement toutes les deux semaines), hormis pendant une courte période de ralentissement au cours du mois de septembre.

2.2 AUX ÉCHELLES RÉGIONALES ET DÉPARTEMENTALES

2.2.1 Région Ile-de-France

La déclinaison de la même approche au niveau des régions et des départements fournit des informations complémentaires. Tout d'abord, le poids de la région Ile-de-France dans les données nationales est très important. Pendant la première vague épidémique de mars-avril, près de la moitié des hospitalisations nationales ont été enregistrées en Ile-de-France (qui compte moins de 20% de la population).

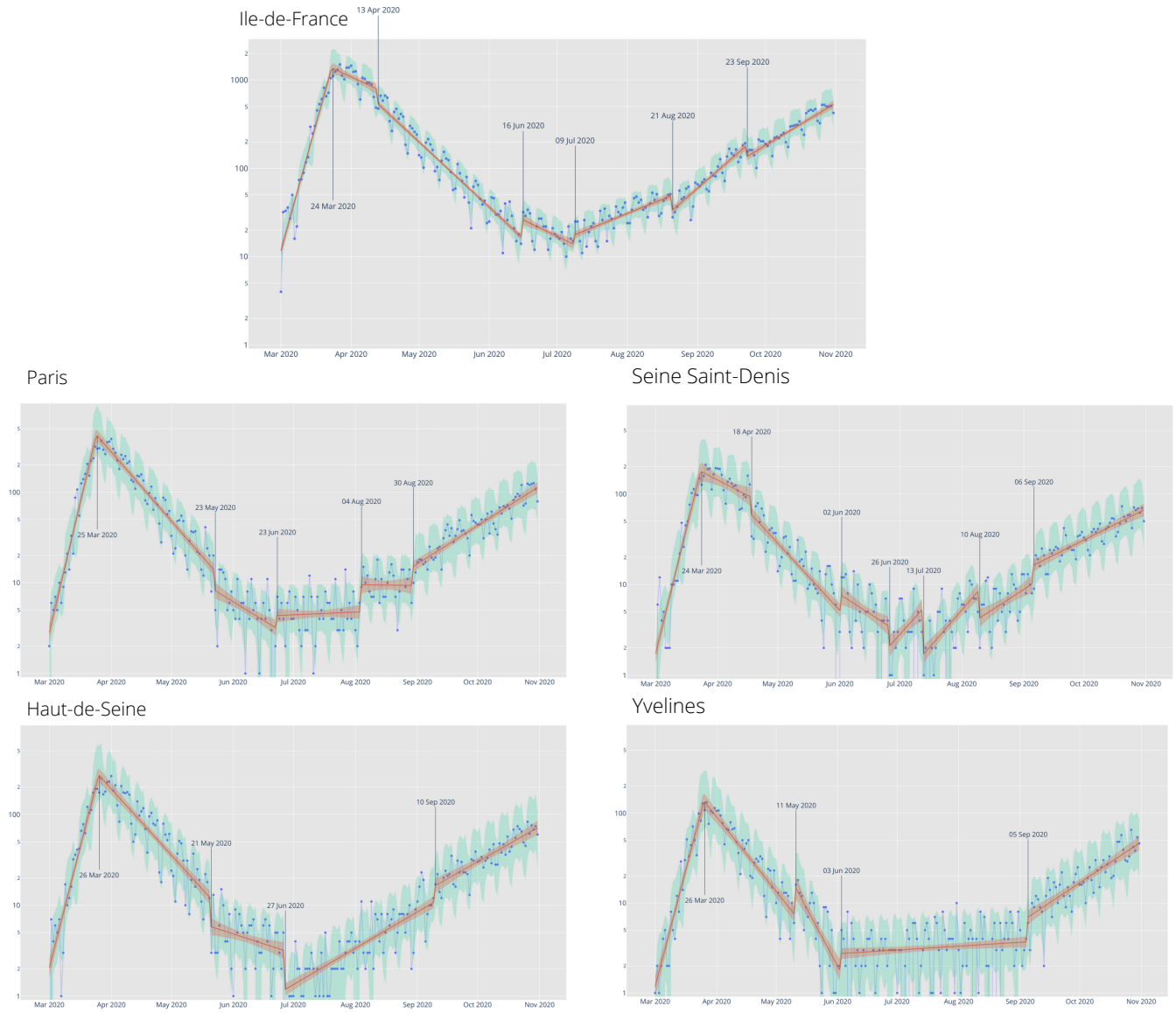


Figure 2.2 Modélisation des données d'Ile de France

La région Ile-de-France se distingue également par une grande homogénéité, avec des départements qui ont tous été assez largement impactés par l'épidémie. On constate néanmoins des différences entre les départements d'Ile-de-France, avec un cumul régional qui agrège des comportements distincts, notamment dans la période qui correspond au confinement, du 25 mars à la mi-juin. Au-delà de ces différences, l'ensemble des courbes départementales présente un minimum dans la période qui va de la dernière semaine de juin à la première quinzaine de juillet. Enfin, à partir de la première quinzaine de septembre on constate dans tous les départements de la région Ile-de-France une augmentation régulière à un rythme soutenu, correspondant à un doublement du nombre d'incidences hospitalières tous les 20 jours. Cette croissance est compatible avec la tendance observée nationalement à partir de la seconde moitié du mois d'août mais sans présenter néanmoins de ralentissement mesurable au cours du mois de septembre.

2.2.2 Autres régions métropolitaines

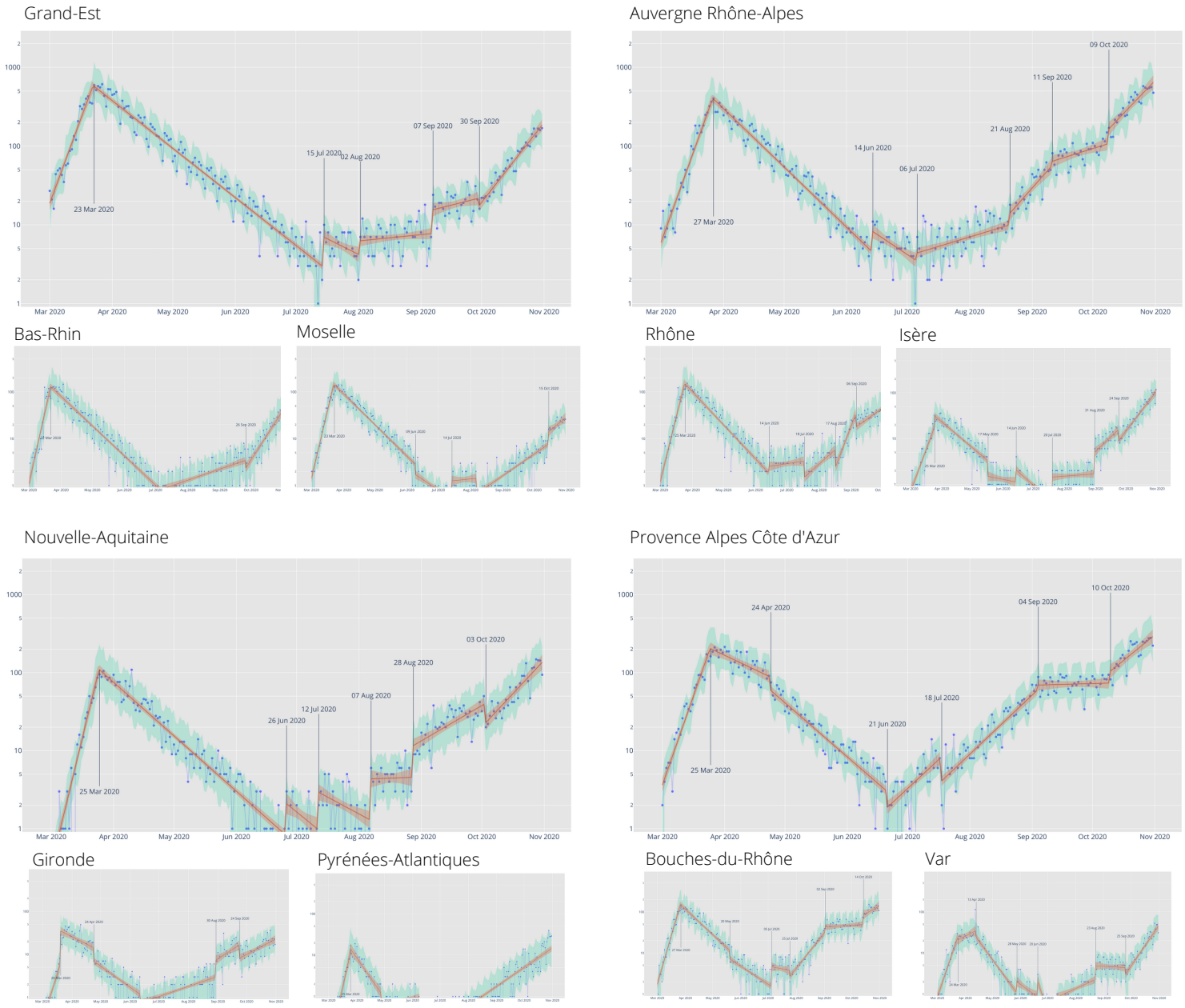


Figure 2.3 Résultats sur quelques régions et départements les plus impactés

La figure 2.3 qui présente les résultats obtenus sur une sélection de régions et départements parmi les plus impactés par l'épidémie montre des situations très différentes. Tout d'abord, on observe au niveau de chaque région des situations beaucoup plus contrastées. Avec des métropoles régionales, comme Lyon et le département du Rhône, par exemple, qui regroupent, pendant la première vague de l'épidémie, près de la moitié des cas enregistrés dans la région Auvergne-Rhône-Alpes (qui comporte pourtant douze départements). La situation est encore plus frappante en Provence-Alpes-Côtes d'Azur où, pendant la première vague, plus de 80% des cas enregistrés l'ont été dans le seul département des Bouches-du-Rhône. Cette constatation implique, à l'inverse, que dans de nombreux départements, le nombre d'hospitalisations est resté faible, devenant même difficilement mesurable de façon fiable pendant les mois de mai à juillet (voir par exemple le cas des Pyrénées-Atlantiques ci-dessus). Néanmoins, on constate rétrospectivement que tous les départements présentent des augmentations du nombre de cas, souvent perceptibles dès le début du mois de juillet et en tout cas détectées sans ambiguïté à partir de la seconde moitié du mois d'août.

A la différence de ce qui est observé dans le Grand Est et en Ile-de-France, on constate par ailleurs que dans de nombreuses régions et départements le niveau atteint à la fin du mois d'octobre est supérieur à celui du pic de la première vague de l'épidémie, confirmant un déplacement important des zones fortement impactées par l'épidémie. Enfin, la période de ralentissement observée durant la seconde quinzaine de septembre sur la courbe nationale ne semble être présente que dans quelques cas : elle est nettement visible sur les courbes des régions telles que Auvergne-Rhône-Alpes et Provence-Alpes-Côte d'Azur (elle est d'ailleurs particulièrement marquée dans certains départements de cette dernière, comme les Bouches-du-Rhône ou le Var).

Table 2.2. Chronologie d'événements en lien avec l'épidémie de Covid-19.

- 12 mars 2020 : Fermeture des crèches, écoles, collèges, lycées et universités
- 17 mars 2020 : Confinement
- 11 mai 2020 : Phase 1 du déconfinement (notamment : réouverture de certains commerces)
- 2 juin 2020 : Phase 2 du déconfinement (notamment : fin de la limite de 100 km pour les déplacements)
- 22 juin 2020 : Phase 3 du déconfinement (notamment : réouverture des écoles et collèges et des restaurants en Ile-de-France)
- 20 juillet 2020 : Le port du masque "grand public" est rendu obligatoire dans tous les lieux clos
- 27 août 2020 : Placement de 19 nouveaux département en zone de circulation active du virus
- 11 septembre 2020 : Passage de 42 département en circulation active du virus
- 05 octobre 2020 : Passage de Paris et de trois départements de la petite couronne en alerte maximale
- 17 octobre 2020 : Début du couvre-feu en Ile-de-France ainsi qu'à Grenoble, Lille, Lyon, Aix Marseille, Saint-Etienne, Rouen, Montpellier et Toulouse
- 24 octobre 2020 : Extension des mesures de couvre-feu à 38 nouveaux départements
- 30 octobre 2020 : Second confinement

Dans la suite de ce document, on présente des observations permettant de caractériser plus systématiquement les phénomènes qui viennent d'être décrits en travaillant

à l'échelle des départements; dans la mesure où on a observé ci-dessus que les données régionales agrègent souvent des comportements très différents;
en utilisant comme référence temporelle les sept périodes définies lors de l'analyse de la courbe nationale qui fournissent un référentiel commun pertinent.

Pour ce faire, on fait appel à d'autres données, permettant de mettre en évidence des phénomènes qui se sont produits pendant l'été, entre les deux phases de l'épidémie.

3. ÉVOLUTION DE LA DENSITÉ DE POPULATION

Au regard de l'analyse de l'évolution de la situation sanitaire, nous pouvons faire deux constats :

1. La période estivale représente un tournant dans le déclenchement de la seconde vague épidémique.
2. La situation sanitaire nationale agrège des réalités très différentes sur le plan régional et départemental.

Afin d'approfondir notre étude, nous examinons les effets de la période estivale sur la redistribution de la population française en nous basant sur une analyse des données de recensement et de mobilité mises à disposition par Facebook dans le contexte de leur programme "*Data for Good*" (cf. Rapport Face au Virus #1 du 25 mai 2020 et Annexe 1 pour une présentation générale de ces données).

3.1 REDISTRIBUTION DE LA POPULATION EN ÉTÉ

3.1.1 Évolution régionale de la population pendant l'été

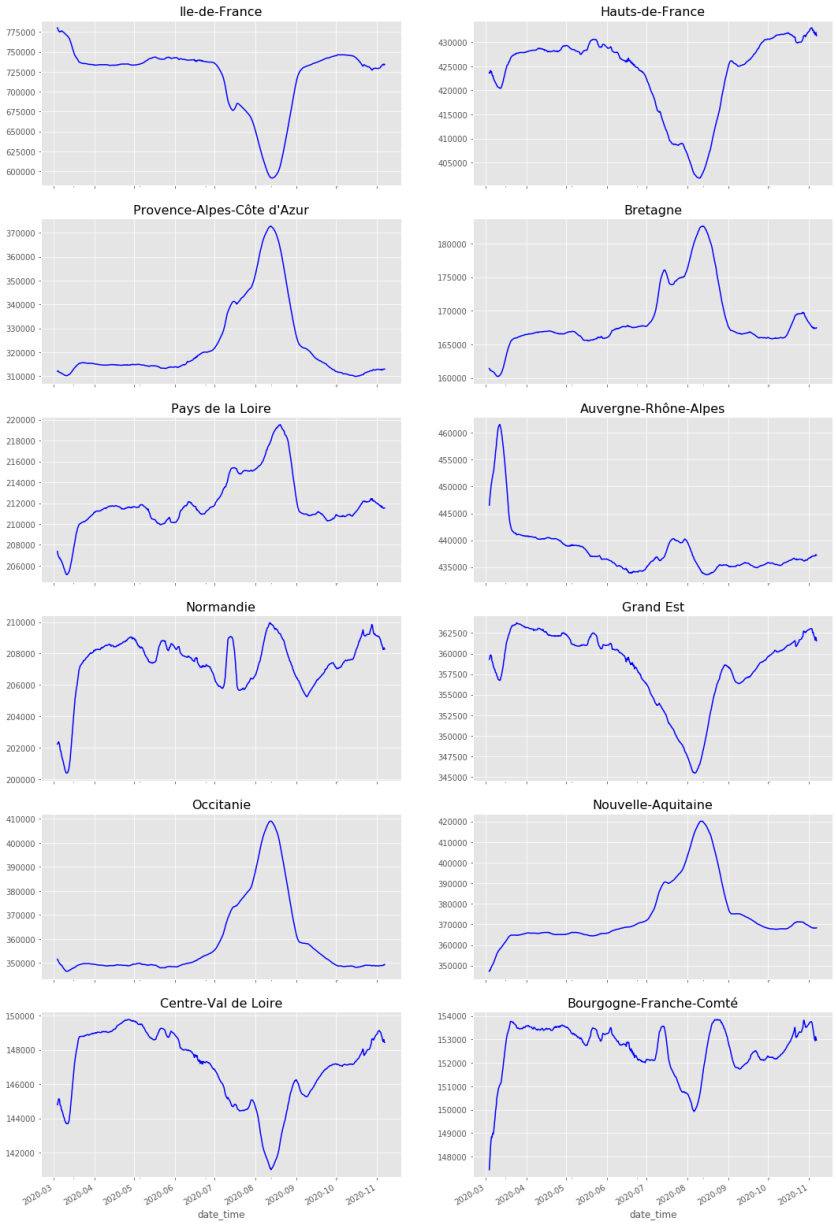


Figure 3.1 Évolution du nombre d'utilisateurs Facebook par région

Les courbes de la figure 3.1 présentent l'évolution du nombre moyen d'utilisateurs Facebook pour chaque jour de la période allant du 06 mars 2020 au 11 novembre 2020 et pour chaque région métropolitaine. **Ces courbes confirment un déplacement important de la population des régions du nord (Ile-de-France, Haut-de-France, Grand Est) vers des régions côtières (Provence-Alpes-Côte-D'Azur, Bretagne, Nouvelle-Aquitaine, Occitanie) pendant la période estivale.**

Les déplacements atteignent leur pic vers mi-août, où l'on enregistre une chute d'environ 2 millions d'utilisateurs Facebook en Île-de-France en parallèle d'une hausse d'environ 60.000 utilisateurs dans la région PACA et 20.000 en Bretagne.

Ces déplacements conduisent non seulement à des mélanges entre populations venant de différentes régions, mais aussi à d'importantes modifications dans la distribution de la densité spatiale de la population comme nous l'illustrons dans la section suivante.

3.1.2 Changements dans la distribution spatiale de la population

IRD : Indicateur de Redistribution de Densité
Les données mises à disposition par Facebook permettent d'estimer le nombre quotidien d'utilisateurs par carreau de surface 600x600 m² en France. La surface sur laquelle la population est estimée étant la même pour chaque carreau, le nombre d'utilisateurs peut facilement être assimilé à une densité de population.

Si on considère l'ensemble des carreaux, il est possible de construire un histogramme, pour chaque jour, qui représente la distribution des densités de population en France à une date donnée.

De façon analogue, il est possible de construire des histogrammes (un par jour) pour représenter les variations relatives de la densité, à l'échelle des carreaux, par rapport à une date de référence. Plus l'écart-type de cet histogramme des variations relatives est grand, plus la distribution spatiale de la densité s'écarte de la population de référence. À l'inverse, quand cette valeur est proche de 0, la distribution est proche de celle à la date référence.

Nous pouvons donc caractériser la redistribution spatiale de la densité de population en France (par rapport à la distribution de référence) **à l'aide de l'écart-type des histogrammes de variation relative de la densité**. Pour chaque jour, on appelle **Indicateur de Redistribution de Densité (IRD)** la valeur de cet écart-type.

Afin de quantifier la redistribution spatiale de la population et de sa densité, nous avons construit un Indicateur de Redistribution de Densité (voir définition dans l'encadré et en annexe) qui quantifie à quel point la distribution de la densité spatiale observée en une journée diffère par rapport à une distribution de référence, que nous avons prise au dimanche 22 mars 2020. La date du 22 mars (premier dimanche suite à l'annonce du confinement) correspond à une situation de référence très atypique dans laquelle la quasi totalité de la population se trouvait à domicile.

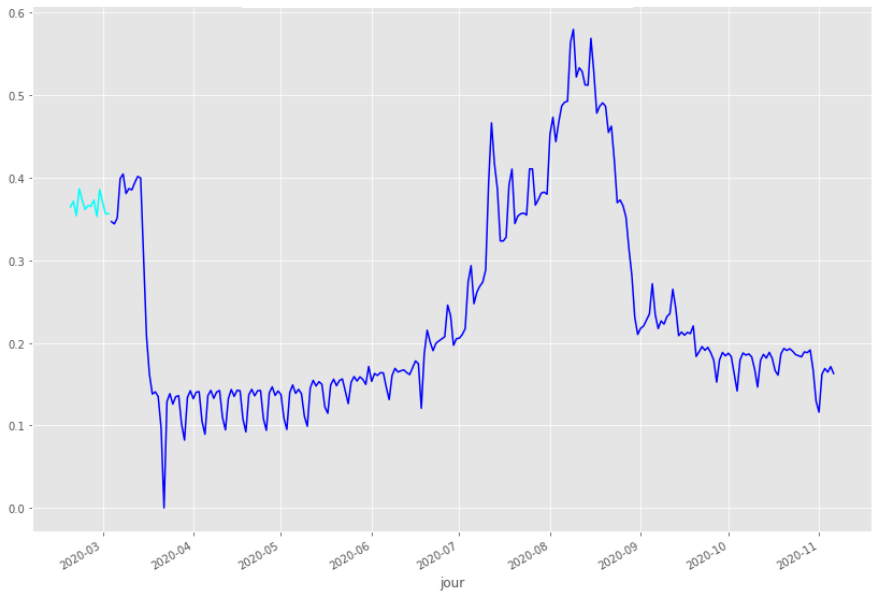


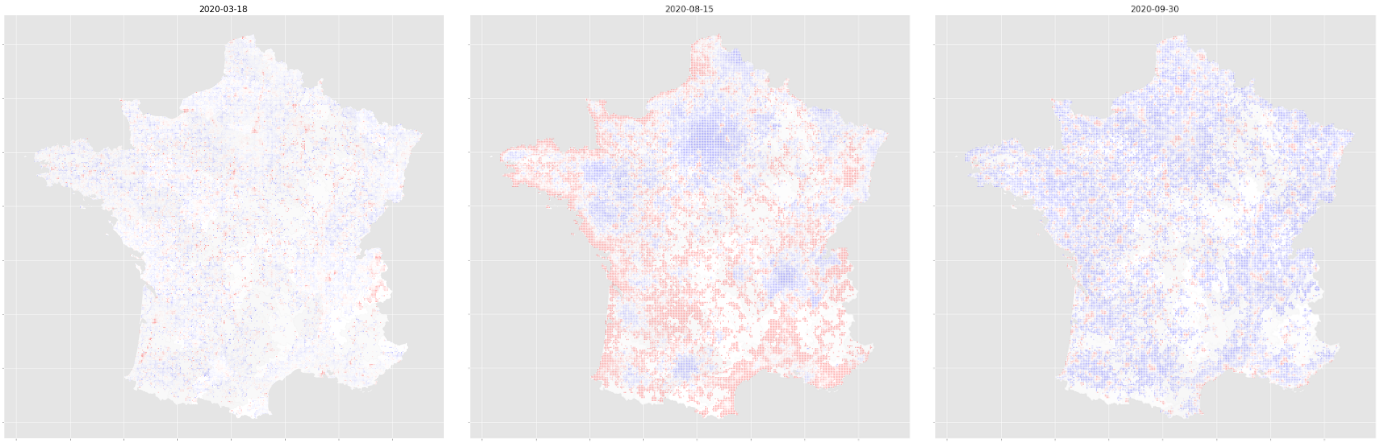
Figure 3.2 Redistribution de la densité des populations à l'échelle nationale

La courbe présentée en figure 3.2 illustre l'évolution de l'indicateur IRD à l'échelle nationale. On observe distinctement une augmentation très significative de l'indicateur aux alentours de la période estivale, avec un pic atteint mi-août. Pendant cette période, la valeur de l'indicateur est même plus forte que celle observée pendant la période pré-confinement de février (en bleu clair). Ceci nous permet de confirmer qu'une **importante redistribution spatiale de la densité de population a eu lieu pendant la période estivale.**

3.2 DENSIFICATION DES ZONES LITTORALES PENDANT L'ÉTÉ

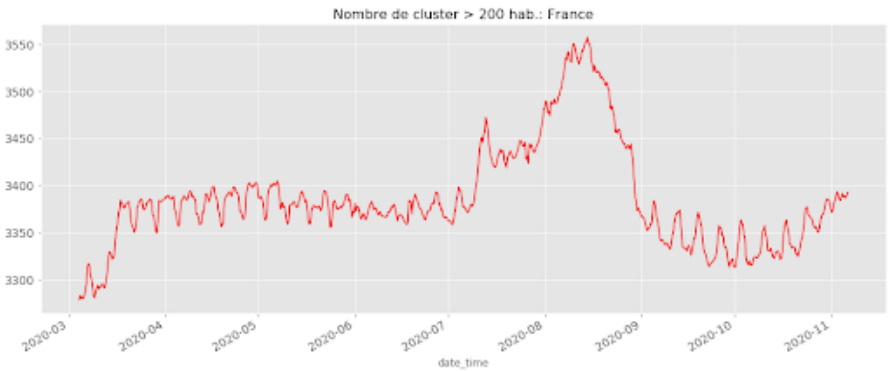
3.2.1 Augmentation de la densité dans de nombreux endroits en été

Figure 3.3 Cartes de variations de densité



La Figure 3.3 illustre la redistribution de la densité à trois dates : une pendant la période de confinement (le 18 mars 2020), une au pic des déplacements estivaux (le 15 août 2020) et une après l'été (le 30 septembre 2020). Les zones en rouge (respectivement en bleu) représentent les carreaux pour lesquels la densité de la population a augmenté (respectivement diminuée) de plus de 25% par rapport à la date de référence du 22 mars. Ces cartes nous permettent de confirmer que **suite aux déplacements estivaux, une partie importante du territoire a vu sa densité augmenter de plus de 25% par rapport à la date de référence.**

3.2.2 Dépeuplement des villes au profit des zones littorales



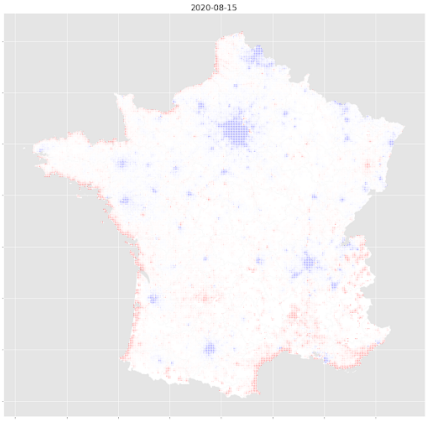


Figure 3.4: Evolution du nombre de carreaux avec plus de 200 utilisateurs en France

La densification de nombreux carreaux à travers le pays a également eu pour effet de faire augmenter en août le nombre de zones très densément peuplées. La figure 3.4 (à gauche) illustre l'évolution du nombre de carreaux comportant plus de 200 utilisateurs, ce qui correspond à des lieux ayant une densité de l'ordre de 9000 habitants/km². Ces valeurs de densité sont très élevées et ne sont habituellement observées que dans les communes les plus densément peuplées de France. Pour mieux visualiser cette augmentation des zones de forte densité pendant le mois d'août, la partie droite de la figure représente les zones de plus de 200 utilisateurs pour lesquelles la densité de la population a augmenté ou diminué de plus de 25% (respectivement en rouge et bleu) pendant la période d'août par rapport à la date de référence du 22 mars. On observe distinctement que **les zones de forte densité habituelle (grandes agglomérations) se sont dépeuplées au profit de certaines zones littorales, fortement prisées pendant l'été.**

4. REGARDS CROISÉS ENTRE LES DONNÉES SANITAIRES ET LES DONNÉS DE DENSITÉ

Nous avons observé dans la partie précédente que, durant l'été, des zones densément peuplées se sont créées en dehors des métropoles où elles se concentrent habituellement. La question posée est de savoir si la présence de zones denses peut être un facteur d'accélération de la propagation du virus, au-delà bien sûr du simple effet de proportionnalité entre le nombre de cas et la taille de la population.

Pour étudier cette question, nous avons mis en correspondance les facteurs de propagation du virus, calculés à partir de la modélisation présentée en Section 2, avec les densités par carreaux fournies par les données Facebook. Nous présentons les sept cartes correspondant respectivement aux différentes périodes déterminées à partir de la courbe nationale d'hospitalisations (cf. Table 2.1).

Lecture des cartes

Sur chaque carte, on représente en vert les carreaux dont la population dépasse 100 utilisateurs Facebook (ce qui correspond à des lieux ayant une densité proche de 4500 habitants/km²) sur chacune des sept périodes identifiées à la section 2. Pour chaque période, deux cartes permettant de qualifier la propagation de l'épidémie au niveau de chaque département sont représentées.
La première (à gauche) correspond au nombre moyen d'hospitalisations par semaine, pour 100 000 habitants, et par département. Plus l'échelle de couleur est foncée, plus le nombre d'hospitalisations rapporté au nombre d'habitants est fort.
La seconde carte (à droite) illustre la valeur du taux de croissance du nombre d'hospitalisations par département estimé selon l'approche présentée dans la Section 2. Les valeurs vont du bleu pour les valeurs négatives (c'est à dire de décroissance de l'épidémie) au rouge pour les valeurs positives, correspondant à une croissance plus ou moins forte. Pour chaque période, compte tenu des incertitudes liées à l'estimation des taux de croissance, on représente sur la carte de droite uniquement trois niveaux correspondant, respectivement, au comportement moyen et aux départements pour lesquels la croissance est, soit, significativement plus faible, soit, plus forte que la moyenne.
La colonne de gauche présente donc le niveau atteint par l'épidémie sur une période donnée, tandis que celle de droite correspond à la tendance de l'évolution sur la même période, fournissant deux informations complémentaires sur la dynamique de l'épidémie. Avec ce choix de visualisation, si les conditions de propagation du virus étaient totalement homogènes sur l'ensemble du territoire, les deux types de cartes devraient être systématiquement monochromes. Ces cartes permettent donc, pour chaque période, de visualiser rapidement l'écart à cette situation de référence.

4.1 PREMIÈRE VAGUE DE L'ÉPIDÉMIE

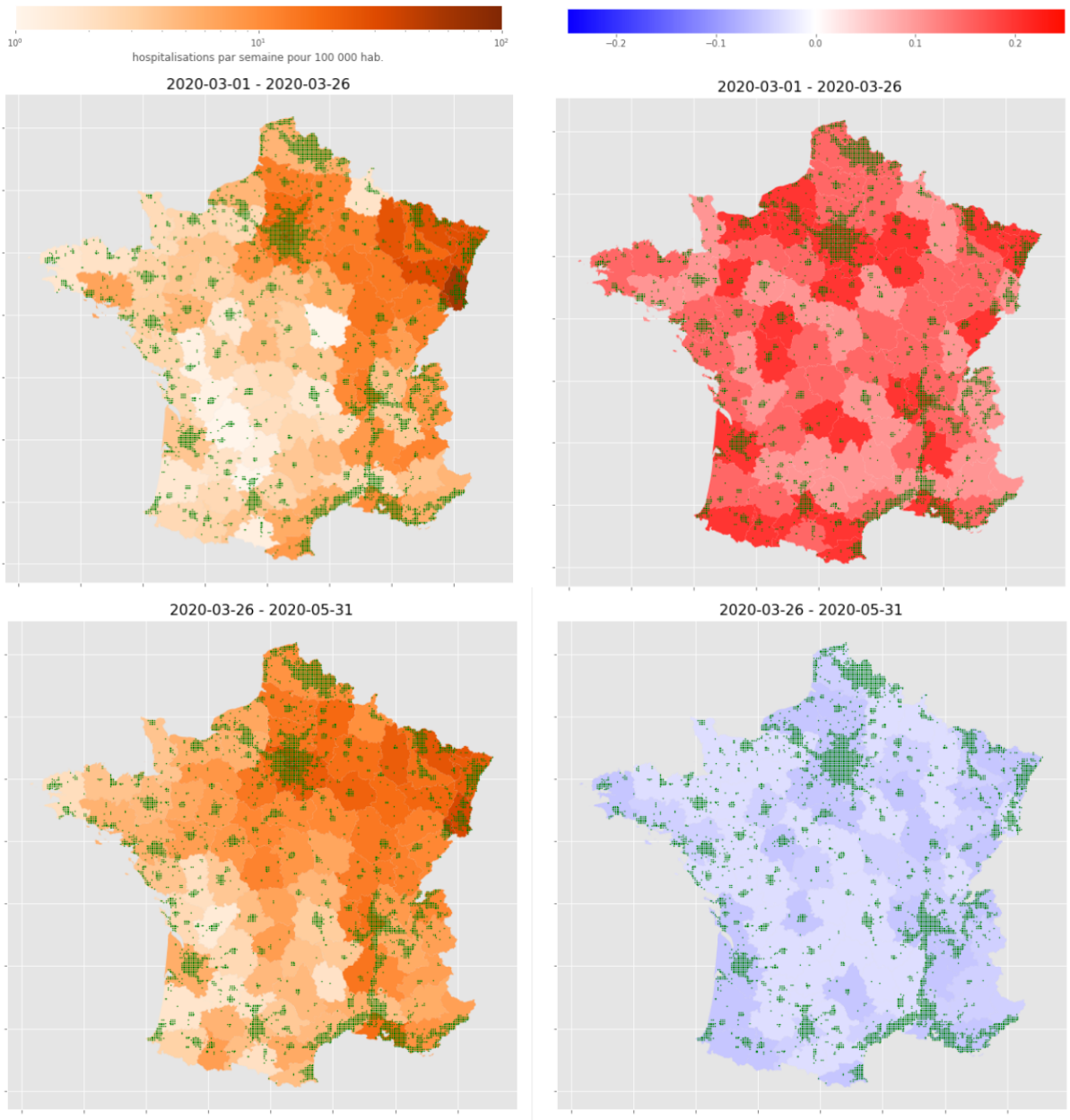


Figure 4.1 Mise en correspondance des données de densité et d'hospitalisation durant la première vague de l'épidémie

Les cartes correspondant à la première vague de l'épidémie (périodes 1 et 2, cf figure 4.1) soulignent les liens qui existent entre la présence de zones à fortes densité de population et la propagation du virus. Ces cartes confirment tout d'abord, qu'en dehors du cas particulier de l'Ile-de-France, il n'y a pas d'homogénéité des situations au sein des régions, ni même entre départements géographiquement voisins. Exception faite de la métropole lilloise qui semble avoir un comportement différent, on constate par ailleurs que la propagation de l'épidémie est plus rapide dans les départements présentant des zones fortement denses.

Cet effet est perceptible en niveau (sur les cartes de gauche) mais surtout notable pour ce qui concerne la vitesse de croissance de l'épidémie (cartes de droite) avec une progression plus rapide (teinte de rouge vif) avant que l'effet du confinement soit mesurable et une décroissance également plus rapide (représenté en bleu plus foncé) pendant la seconde période correspond au confinement. Il est important de rappeler que, sur la carte de gauche (et contrairement aux visualisations les plus courantes de l'épidémie de Covid-19), les nombres d'hospitalisations ont été normalisés par la population des départements. Par conséquent, la corrélation avec la présence de zone denses n'est pas due au simple fait que celles-ci correspondent également aux départements les plus peuplés. De même, le taux de croissance de l'épidémie (cartes de droite) n'a a priori aucun lien avec la population, on constate d'ailleurs avant confinement des progressions rapides dans des départements peu peuplés comme la Mayenne ou la Corrèze. Compte tenu de ces différentes observations, la présence de zones de forte densité apparaît bien comme facteur explicatif de l'accélération de la propagation de la pandémie durant la première vague.

Les cartes de la figure 4.1 confirment également les observations bien établies concernant la diffusion de l'épidémie dans le territoire français durant la première vague : une épidémie qui débute dans la région de l'Est, puis dans le nord de l'Ile-de-France, avant de se propager le long des grands axes de communication, notamment vers la région Lyonnaise et le Sud-Est. Globalement, l'Ouest de la France a été largement épargné par la première vague épidémique. Il convient néanmoins de tenir compte du fait que le confinement a interrompu la progression de l'épidémie et au vu du taux de croissance mesuré pendant la première période du mois de mars, par exemple, en Gironde, on peut imaginer que cette constatation aurait été remise en cause si le confinement avait été mis en place plus tardivement.

4.2 TRANSITION VERS LA PÉRIODE ESTIVALE

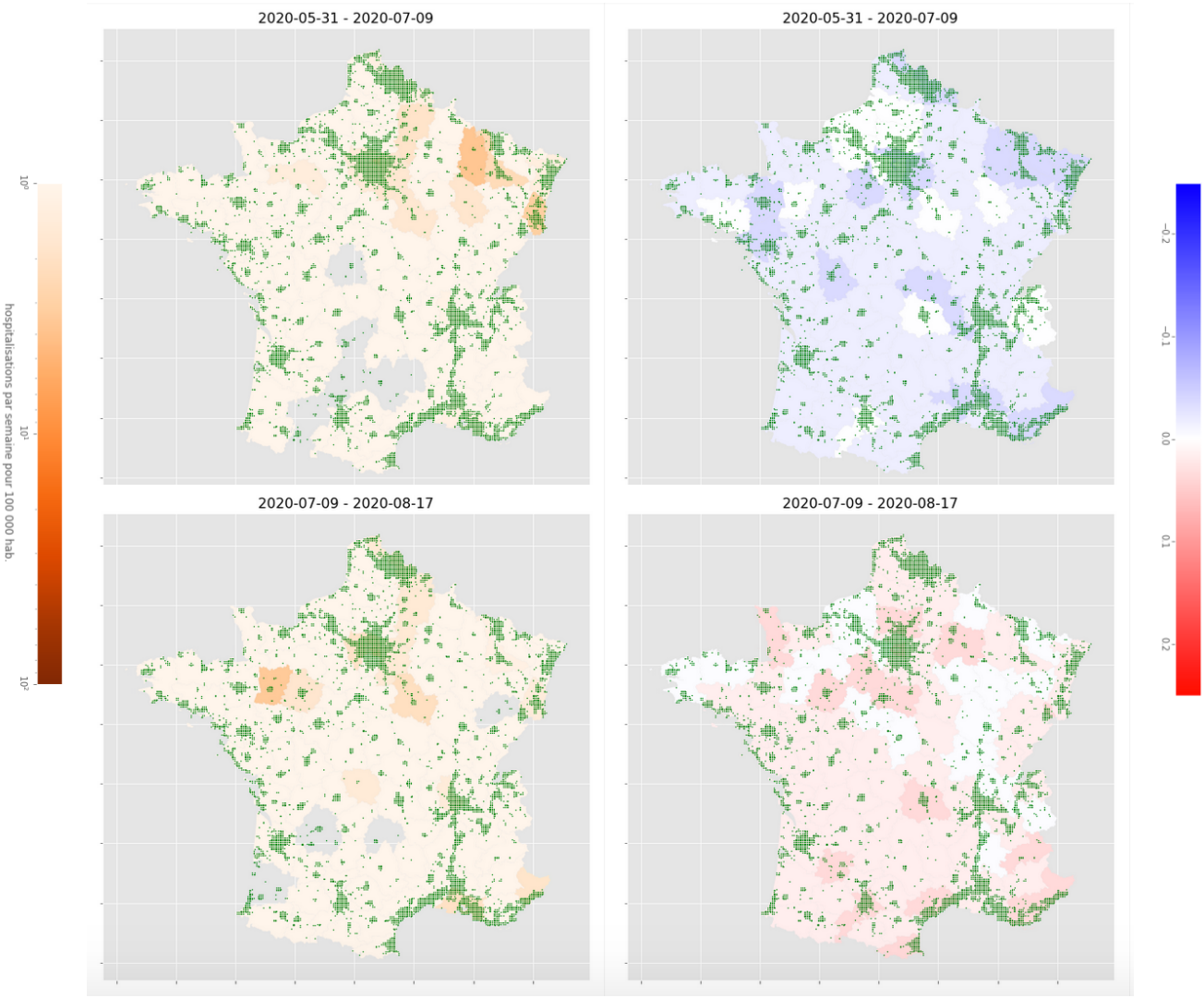


Figure 4.2 Mise en correspondance des données de densité et d'hospitalisation durant l'été

Les deux périodes suivantes correspondent à une situation transitoire qui va, compte tenu du délai de quinze jours nécessaire pour que les phénomènes soient perceptibles sur les données d'hospitalisation, de la fin du confinement national (11 mai) jusqu'au mois d'août. Durant cette phase, le nombre de cas baisse, jusqu'à atteindre des niveaux difficilement mesurables en utilisant les données d'hospitalisations. On constate par ailleurs que les régions où l'impact de l'épidémie a été très fort pendant la première vague épidémique, en particulier le Grand-Est, ne se différencient plus fortement des autres régions. Néanmoins, les cartes qui présentent les taux de croissance (à droite) montrent clairement que cette phase transitoire se décompose en deux périodes successives distinctes. Dans la première, les nombres d'hospitalisations continuent à décroître mais plus lentement, voire stagnent dans certains départements comme celui de Paris. A l'inverse, **dans la seconde période (à partir du 7 juillet) les nombres d'hospitalisations se remettent à croître légèrement dans la plupart des départements, avec des croissances observées dans des départements souvent peu impactés jusqu'à présent par l'épidémie, comme l'Hérault, les Alpes Maritimes ou les Hautes Alpes.**

4.3 REPRISE DE L'ÉPIDÉMIE

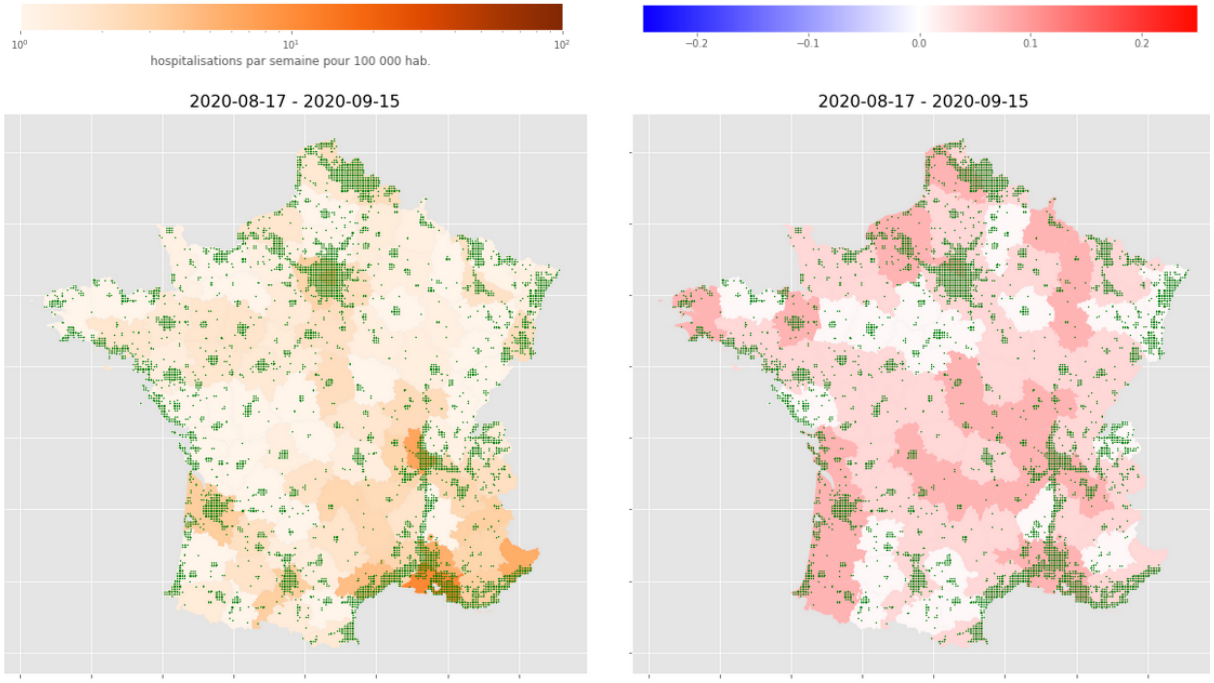


Figure 4.3 Mise en correspondance des données de densité et d'hospitalisation durant l'été

La date de la mi-août marque un net tournant, visible sur la carte présentant le nombre de cas (à gauche) pour la période du 17 août au 15 septembre. Plusieurs départements retrouvent des niveaux d'hospitalisations très significatifs (de l'ordre de dix hospitalisations journalières pour 100 000 habitants), dont en particulier les Bouches-du-Rhône. De façon concomitante, **la localisation de l'épidémie apparaît sous une forme différente de ce qu'elle était dans la première vague, avec des départements fortement impactés qui se concentrent principalement dans le Sud-Est.** Les facteurs de croissance correspondant à cette première période sont également significatifs sur les départements du littoral ouest ainsi que dans plusieurs départements ne comportant pas ou peu de zones fortement denses, notamment en région Auvergne-Rhône-Alpes.

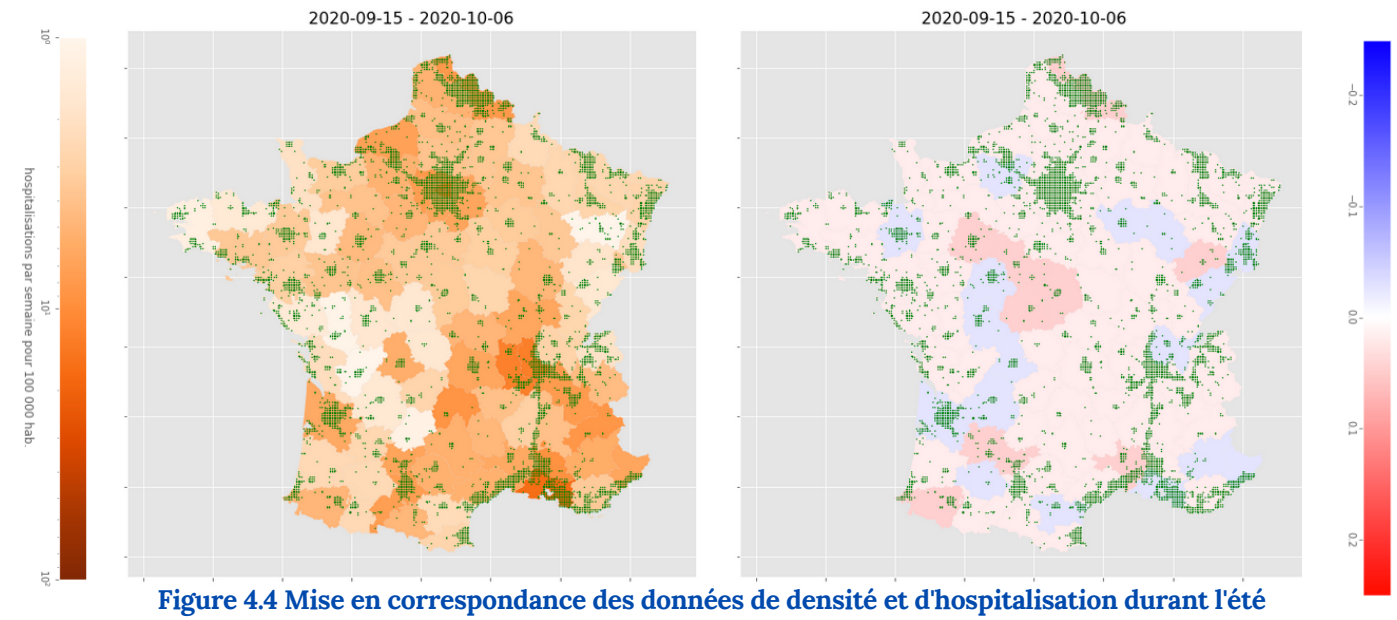


Figure 4.4 Mise en correspondance des données de densité et d'hospitalisation durant l'été

Cette première période est suivie par un épisode relativement court dans lequel, à la fois, l'épidémie s'est largement diffusée sur le territoire, mais où elle également a pu localement donner l'impression de ralentir, comme le signalent sur la carte de droite les départements en bleu pour lesquels le nombre d'hospitalisations journalières est globalement en décroissance sur la période. De façon intéressante, une bonne partie des départements concernés étaient au contraire en croissance forte dans la période précédente (Eure, Ile-et-Vilaine, Gironde, Ain, Haute Vienne et surtout surtout Bouches-du-Rhône où le phénomène est particulièrement marqué). Sur la carte de gauche qui indique le niveau de l'épidémie, on constate que, sur la même période, les grandes métropoles du Nord de la France (Paris, Lille) reviennent à des nombre de cas (rapportés à la population) significatifs. **L'ensemble de ces observations suggèrent que la relative accalmie observée durant cette période est principalement dûe à un effet de redistribution des populations sur le territoire suite à la fin de la période estivale à partir du début du mois de septembre.**

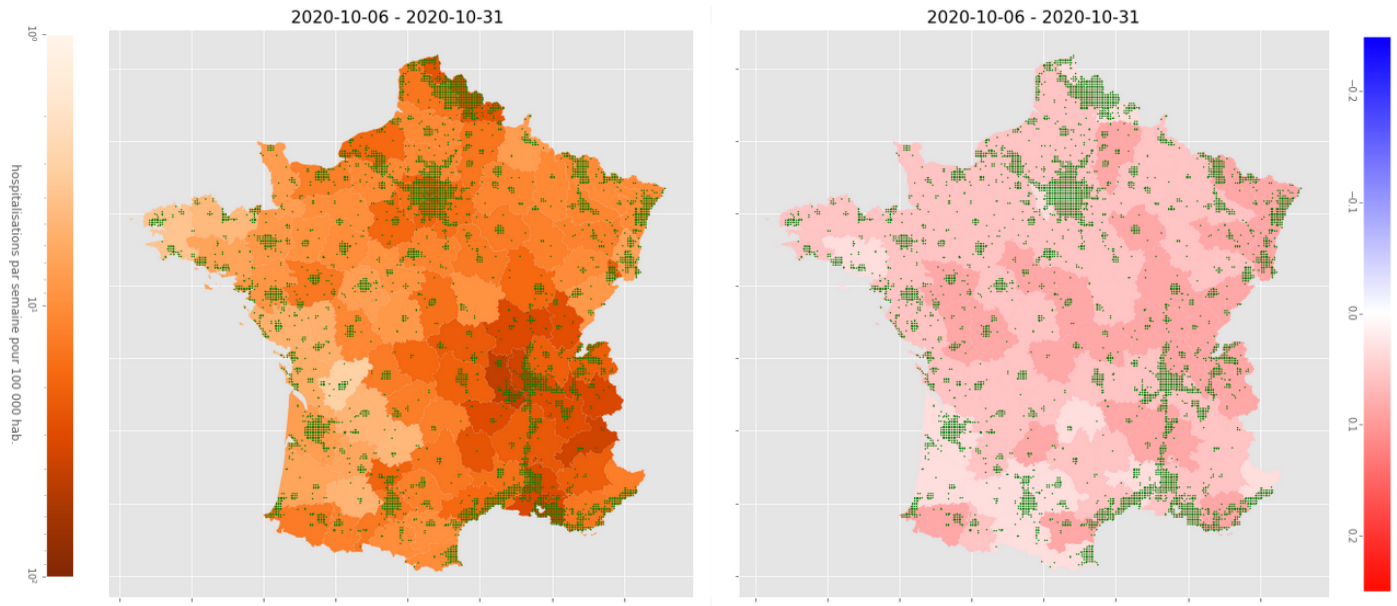


Figure 4.5 Mise en correspondance des données de densité et d'hospitalisation durant l'été

Enfin la dernière période voit l'épidémie reprendre à un rythme de croissance proche de celui du mois d'août, avec un impact particulièrement significatif dans les départements où la croissance de l'épidémie a été forte pendant l'été, c'est-à-dire principalement dans le quart Sud-Est de la France. **A la différence de la première vague, on observe également que les niveaux les plus significatifs (en terme de nombre de cas rapportés à la population) sont souvent atteints dans des départements faiblement peuplés et comportant peu de zones denses (Savoie, Hautes-Alpes, Loire, Saöne-et-Loire, Jura).**

4.4 DENSITÉ ET TAUX D'HOSPITALISATIONS PAR HABITANT

L'analyse visuelle des cartes précédentes suggère un lien entre l'intensité de l'épidémie et la densité de population dans le sens où il semble que des **zones à haute densité figurent souvent parmi les zones ayant les plus forts taux d'hospitalisation par habitant**.
Afin d'étudier ce point, pour chaque jour nous avons listé par ordre décroissant les départements par rapport à leur nombre d'hospitalisations par habitant. Nous avons ensuite compté la fréquence d'apparition de chaque département dans le top 20% de cette liste pour chacune des sept périodes considérées. Ce calcul permet donc d'estimer quels sont les départements qui sont le plus fréquemment touchés à chaque période.

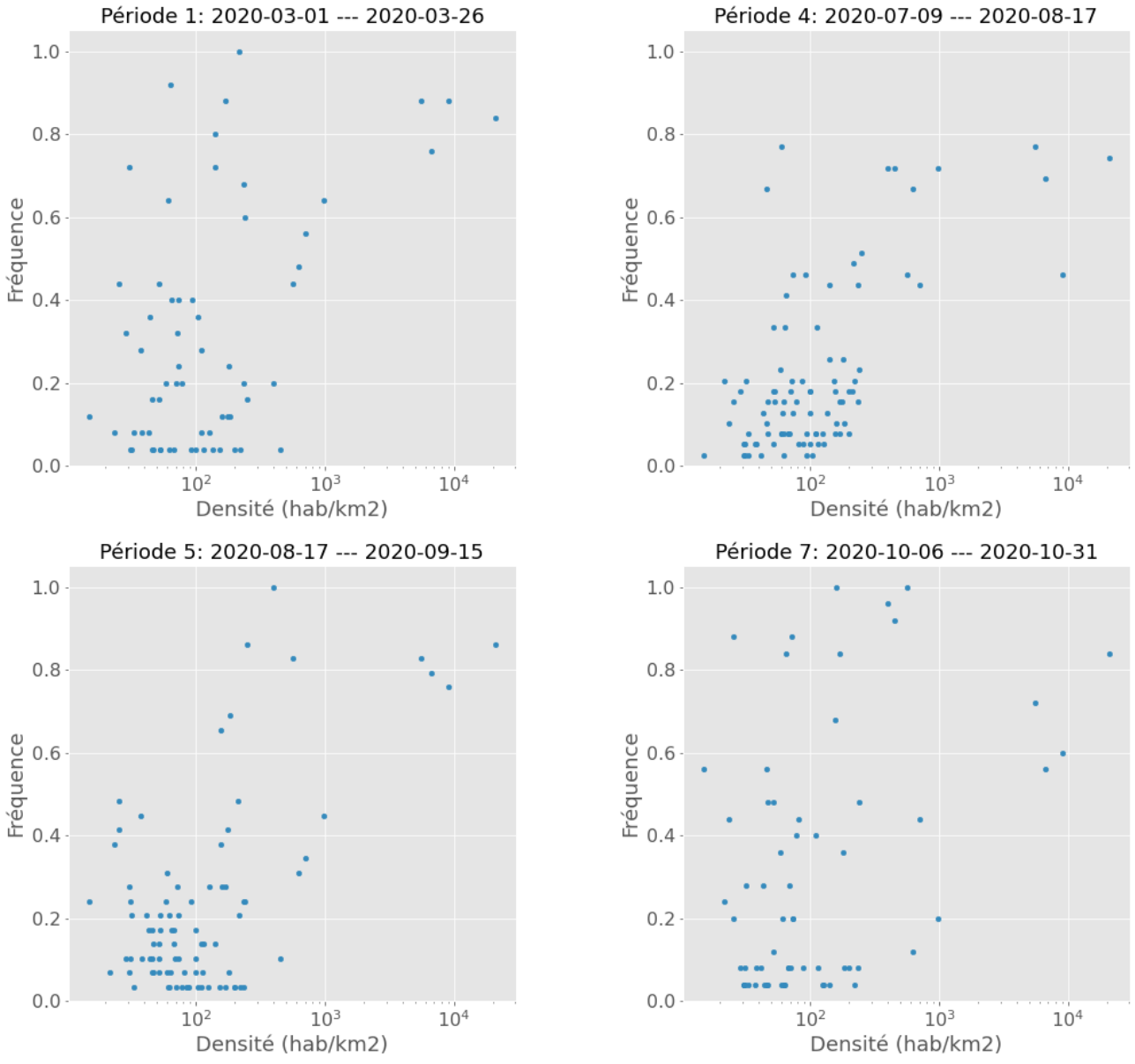


Figure 4.6 : Pour quatre périodes, fréquence d'apparition dans le top 20% des départements avec le plus d'hospitalisations par habitant

Les figures 4.6 illustrent cette fréquence d'apparition en fonction de leur densité de population. Lorsque la fréquence vaut 1, alors le département en question a figuré tous les jours dans le top 20% des départements avec les plus fortes hospitalisations par habitant. À l'inverse, une fréquence de 0 signifie que le département n'a jamais été parmi les 20% des départements les plus touchés.

Il apparaît assez clairement que **les départements à forte densité figurent très souvent parmi les départements avec le plus d'hospitalisations par habitant.** Il est aussi à noter que **certains départements à densité modérée présentent parfois des fréquences élevées d'apparition mais seulement sur certaines périodes**. Ceci s'explique par des facteurs ou des événements ayant eu lieu dans chaque période. Par exemple, dans la première vague, la direction claire de propagation de l'épidémie du nord-est vers l'ouest explique pourquoi certains départements du nord-ouest à densité modérée figurent parmi les départements les plus souvent atteints.

L'étude de cette même fréquence d'apparition sur l'ensemble de tous les segments révèle cette fois-ci une corrélation forte entre la densité et la fréquence d'apparition dans les départements les plus touchés (voir figure 4.7). Ceci confirme l'observation précédente que **les départements à densité faible ou modérée jouent un rôle plutôt ponctuel pendant certaines périodes, alors que le cœur de l'épidémie semble être localisé dans les zones à forte densité.**

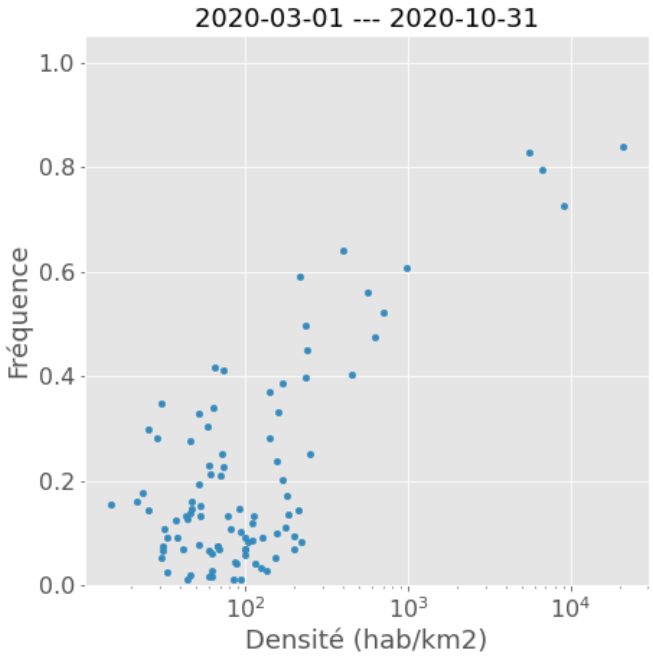


Figure 4.7 : Pour l'ensemble de la période, fréquence d'apparition dans le top 20% des départements avec le plus d'hospitalisations par habitant

L'étude de cette même fréquence d'apparition sur l'ensemble de tous les segments révèle cette fois-ci une corrélation forte entre la densité et la fréquence d'apparition dans les départements les plus touchés (voir figure 4.7). Ceci confirme l'observation précédente que **les départements à densité faible ou modérée jouent un rôle plutôt ponctuel pendant certaines périodes, alors que le cœur de l'épidémie semble être localisé dans les zones à forte densité.**

AUTEURS ET REMERCIEMENTS

JAMAL ATIF (DAUPHINE - PSL, LAMSADE)
BERTRAND CABOT (CNRS, IDRIS)
OLIVIER CAPPÉ (CNRS, DI ENS - PSL)
OLGA MULA (DAUPHINE - PSL, INRIA)
RAFAËL PINOT (DAUPHINE - PSL, LAMSADE)

MERCI À

Laurent Massoulié (Inria, DI ENS), Hadrien Bernard-De-Dompsure (CNRS, IDRIS), la DREES pour les données SI-VIC, et Facebook pour les données de géolocalisation,

ET À

Kevin Lamothe (ESPCI Paris - PSL) pour l'édition et la mise en page.

Contact scientifique : Jamal Atif jamal.atif@dauphine.psl.eu
Contact Communication : communication@psl.eu

Annexe 1: Redressement des données Facebook

Les données utilisées dans ce rapport ont été fournies par Facebook à l'Université PSL, dans le cadre du programme « Data for Good ». Les données de population fournies par Facebook représentent environ 4 millions d'utilisateurs géolocalisés sur toute la France (soit environ 6% de la population française).

Elles ont été collectées et traitées dans l'objectif de permettre aux chercheurs de mieux comprendre la dynamique de la pandémie de COVID-19 et de développer des modèles et des analyses plus précises. Les données utilisées dans ce travail, sont anonymisées et agrégées par rapport au niveau de granularité que nous utilisons (carreau ou département). Pour une description plus détaillée des données ainsi que des mécanismes d'anonymisation appliqués, nous référons le lecteur au descriptif proposé par Facebook [1]. En raison des mécanismes d'anonymisation, du faible nombre d'utilisateurs Facebook dans certaines zones, mais aussi de certains biais de population [2], il convient de rester prudent quant aux conclusions qui peuvent être tirées de l'observation de ces données. Cependant, l'échantillon de la population que représentent les utilisateurs Facebook semble être suffisamment représentatif pour identifier les zones de fortes densités (comme nous l'avons fait dans les sections 3 et 4). Pour illustrer ce propos, nous présentons dans la figure A.1 une comparaison entre une carte de densité de population par commune en France (à gauche) et une visualisation de la densité d'utilisateurs Facebook par carreau (à droite) en date du 03 mai 2020. Visuellement, les deux cartes correspondent très bien, ce qui nous conforte dans l'idée que, au moins à un niveau macroscopique, les cartes de densité de Facebook correspondent à une réalité démographique.

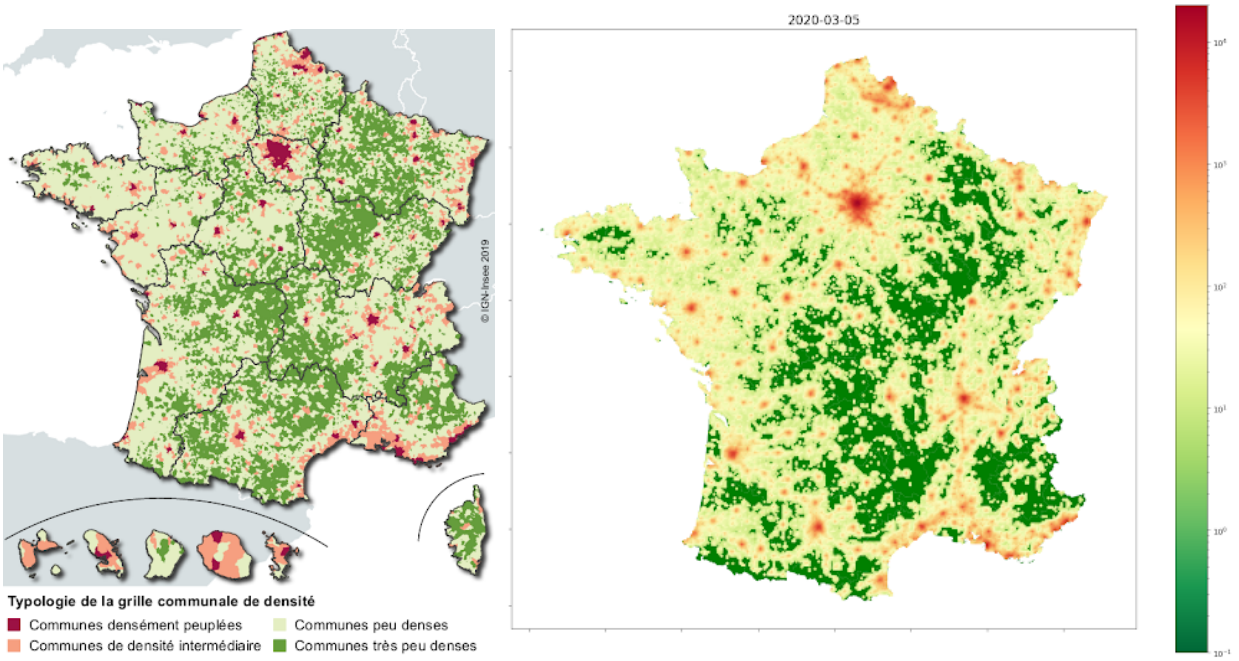


Figure A.1. Cartes de densité de population. À gauche, carte INSEE tirée de [3], à droite carte Facebook.

Redressement simple de la population Facebook:

Pour faire correspondre un nombre d'utilisateurs Facebook à une population réelle, on peut multiplier les utilisateurs par un facteur 16.75 (c'est à dire 67/4). Il est cependant important de remarquer que ce redressement simple ne prend pas en compte les potentiels déséquilibres de représentativité des données, nous utilisons donc ce redressement purement à usage indicatif, notamment dans les sections 3 et 4.

[1] Maas Paige, et al."Facebook Disaster Maps: Aggregate Insights for Crisis Response and Recovery. "Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain. 2019.
[2] https://www.facebook.com/ads/audience-insights/people?act=265465104820229&age=&18-country=FR
[3] https://www.insee.fr/fr/statistiques/4252859

Annexe 2: Modélisation des données d'hospitalisations

Dans le cadre de l'épidémie de Covid-19, trois types de facteurs compliquent l'estimation des paramètres de modèles épidémiologiques. Les premiers sont liés à la nature de la maladie, en particulier, la vitesse de transmission du virus (avec une période contagieuse inférieur à sept jours), la proportion non négligeable de personnes porteuses du virus mais asymptomatiques, donc non détectables, ou la part, majoritaire, de personnes chez qui la maladie se traduit par des symptômes bénins et dont le suivi est très variable. Par conséquent, les informations disponibles sur la propagation de l'épidémie restent partielles et certains aspects clés comme la proportion de personnes asymptotiques, le nombre de personnes infectées par le virus depuis le début de l'épidémie ou la durée de la période contagieuse ne peuvent être connus que de façon indirecte.

Le second facteur est lié à l'ampleur de l'épidémie qui a nécessité la mise en place de multiples mesures exceptionnelles (confinement, restriction de certaines activités, port du masque, etc.) Ces mesures produisent des effets, parfois majeurs, sur la dynamique de l'épidémie qui ne permettent pas d'utiliser un modèle unique qui expliquerait l'ensemble des données observées depuis le début de l'épidémie.

Enfin le troisième facteur est lié à la qualité des données disponibles qui résultent de l'agrégation complexe de sources multiples, renseignées manuellement sur l'ensemble du territoire national. La conséquence la plus manifeste de ce mode de collecte des données est la présence de très forts effets de variations hebdomadaires (qui traduisent essentiellement une forte baisse de l'activité lors des week-ends). Dans la plupart des visualisations de l'épidémie présentées habituellement, ces effets sont masqués, soit en considérant des moyennes glissantes (sur une ou deux semaines) soit par effet de normalisation (par exemple lorsque le taux de positivité des tests est représenté). Néanmoins, lorsque l'on souhaite dépasser la simple visualisation des données et associer une mesure de confiance statistique aux traitements effectués, il est nécessaire de prendre en compte plus précisément la nature des variations observées dans les données.

Pour tenir compte de ces différents éléments, nous avons choisi pour cette étude :

- De n'utiliser que les données d'incidence (ou de nouvelles admissions) hospitalières qui sont consolidées depuis le début du mois de mars 2020 (données dites SI-VIC), en utilisant les séries historiques corrigées mises à disposition par la DREES - Ministère des Solidarités et de la Santé. (voir [3])
- De n'estimer que le facteur de croissance exponentielle local de l'épidémie, qui peut être interprété comme une approximation locale du comportement de modèles épidémiologiques plus complexes. (voir [4])
- D'estimer ce facteur de croissance sur des périodes qui peuvent être considérées comme stables, ces périodes étant déterminées par une approche, classique dans le domaine de l'analyse de séries temporelles, de détection de ruptures par maximisation de la vraisemblance par programmation dynamique.
- D'inclure une contribution périodique hebdomadaire (stable au cours du temps), estimée à partir des données (il est d'ailleurs apparu clairement que cette tendance périodique devait être estimée pour chaque série indépendamment).
- De ne pas utiliser de forme de régularisation ou d'agrégation spatiale pour préserver, autant que possible, les spécificités locales des données (ce qui au vu des résultats apparaît très important). A l'inverse, nous avons considéré, de façon indépendante, les séries de données agrégées aux trois échelles, nationale, régionale et départementale.

Enfin le dernier point clé concerne la modélisation de la variabilité des données d'hospitalisations pour laquelle il est apparu nécessaire de concilier deux caractéristiques importantes : d'une part, une variabilité qui est globalement proportionnelle au nombre de cas (qui se traduit par des variations d'ampleur à peu près constante lorsque les données sont représentées sur une échelle logarithmique) et, d'autre part, la présence de données nulles (correspondant à des jours sans hospitalisation enregistrée), très fréquentes en particulier à l'échelle départementale ainsi que durant la période de creux de l'épidémie (de mai à juillet 2020). Nous utilisons pour ce faire un modèle statistique spécifique dit de *Zero-Inflated Log-Normal Regression* (ZILNR) dans lequel les paramètres sont estimés par minimisation du critère suivant :

$$\sum_t \mathbb{1}_{\{Y(t)=0\}}\lambda(t) + \mathbb{1}_{\{Y(t)>0\}}\left(-\log(1-e^{-\lambda(t)}) + \log(2\pi)/2 + \log(\sigma) + \frac{(\log Y(t) - \log\lambda(t))^2}{2\sigma^2}\right)$$

où $Y(t)$ désigne le nombre d'hospitalisations au temps t, $\log\lambda(t)$ est la tendance estimée en échelle logarithmique et σ correspond à un paramètre de dispersion, homogène à un écart type (en échelle logarithmique). L'expression $\mathbb{1}_C$ est à interpréter comme valant 1 si la condition C est vérifiée et 0 sinon. Ce modèle représente la distribution statistique

des valeurs non nulles, au temps t, par une loi log-normale de paramètres $(\log \lambda(t), \sigma)$, tandis que l'occurrence d'une valeur nulle se fait avec probabilité $e^{-\lambda(t)}$. Cette dernière valeur coïncide avec ce que donnerait une modélisation par une loi de Poisson, qui ne peut cependant pas être utilisée pour l'ensemble des données, au risque de sous-estimer de façon très significative leur variabilité. Pour identifier un taux de croissance exponentielle, on représente $\lambda(t)$ sous la forme

$$\log \lambda(t) = \alpha + \beta t + s(t)$$

où $s(t)$ est une contribution périodique centrée de période 7 jours (donc définie par 6 paramètres). Pour chaque série de données, les paramètres σ et $s(t)$ sont estimés globalement sur l'ensemble de la série, tandis que α et β sont estimés séparément sur chacune des périodes identifiées par la programmation dynamique. Le paramètre central d'intérêt β correspond au taux de croissance exponentielle, qui peut être converti en facteur de croissance à 14 jours, en considérant $e^{14\beta}$. Les intervalles de confiance sur les paramètres α et β estimés sont obtenus via l'approximation asymptotique normale en utilisant la matrice d'information de Fisher liée au modèle. Les intervalles de fluctuations (ou prédictifs) sur les données sont obtenus en additionnant, au temps t, la variance de l'incertitude sur l'estimation de $\lambda(t)$ à la variance résiduelle σ^2 et en calculant les quantiles de la loi marginale dans le modèle ZILNR. Les incertitudes sur l'estimation de σ ne sont pas prises en compte. Tous les intervalles de confiance sont calibrés au niveau de couverture (asymptotique) symétrique de 90% (5% par excès et 5% par défaut). Pour déterminer la segmentation en périodes homogènes, l'algorithme de programmation dynamique a été utilisé en excluant les segments d'une durée inférieure à quinze jours et le nombre de périodes retenues pour la courbe nationale a été déterminé en utilisant le critère BIC (*Bayesian Information Criterion*).

[3] Plateforme Covid-19. URL: https://covid-19.sante.gouv.fr/
[4] Wallinga, J., and M. Lipsitch. "How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers." Proceedings: Biological Sciences 274, no. 1609 (2007): 599-604.

Annexe 3: Indice de Redistribution de Densité

La population française est distribuée de façon très hétérogène dans le territoire, ce qui engendre une distribution spatiale de la densité de population qui est aussi très hétérogène. Cette distribution spatiale évolue au cours du temps à cause des déplacements qui ont lieu, ce qui entraîne que certaines régions peu denses souffrent ponctuellement de fortes densifications (ou réciproquement). L'Indicateur de Redistribution de Densité a été construit pour quantifier à quel point cette distribution spatiale de densité change par rapport à une distribution spatiale de référence. La démarche suivie pour le calculer permet aussi de détecter les zones qui se sont fortement densifiées/dépeuplées. Nous détaillons avec quelques équations comment l'IRD est calculé.

Les données de Facebook nous informent sur le nombre d'utilisateurs et les déplacements sur un découpage du territoire en carreaux de surface $S = 600 \times 600\, m^2$. Afin de garantir l'anonymat, seule l'information sur les carreaux avec un nombre minimal d'utilisateurs nous est transmise. Notons K le nombre de carreaux et notons $n(d,k)$ le nombre de d'utilisateurs Facebook détectés au jour d dans le carreau k pour $k = 1, \dots, K$. La densité d'utilisateurs Facebook est donnée par

$$\mu(d,k) = \frac{n(d,k)}{S}, \quad k = 1, \dots, K.$$

Comme la surface S est identique pour tous les carreaux, étudier la distribution spatiale et temporelle du nombre d'utilisateurs revient à étudier la distribution en densité d'utilisateurs (modulo un facteur multiplicatif). Par ailleurs, comme expliqué en Annexe 1, il est raisonnable de supposer que le nombre d'utilisateurs est proportionnel au nombre réel de personnes se trouvant sur chaque carreau. En admettant ceci, le nombre d'utilisateurs $n(d,k)$ est donc proportionnel à la densité de population $\rho(d,k)$ se trouvant réellement dans chaque carreau k. Par conséquent, nous avons les proportionalités

$$n(d,k) \sim \mu(d,k) \sim \rho(d,k)$$

ce qui veut dire que **nous pouvons étudier les distributions réelles de densité à travers le nombre d'utilisateurs Facebook.**

L'indice requiert de fixer un jour de référence, que nous notons $\bar{d}$. Pour chaque carreau k, nous prenons $\bar{n}(k) = n(\bar{d},k)$ comme le nombre d'utilisateurs de référence. La densité réelle de référence $\bar{\rho}(k) = \rho(\bar{d},k)$ est proportionnelle à $\bar{n}(k)$. Pour un jour d, nous pouvons ainsi calculer la variation relative de densité par rapport au jour de référence comme

$$r(d,k) = \frac{n(d,k) - \bar{n}(k)}{\bar{n}(k)} = \frac{\rho(d,k) - \bar{\rho}(k)}{\bar{\rho}(k)}, \quad k = 1, \dots, K.$$

Cette valeur dit que si $r(d,k)$ vaut une certaine valeur x, alors la densité $\rho(d,k)$ au jour d et au carreau k a augmenté de $x\%$ par rapport à sa valeur de référence $\bar{\rho}(k)$. Notons que $r(d,k)$ peut être négatif, auquel cas la densité aurait diminué de $x\%$ (toujours par rapport à la référence). Notons aussi qu'au jour $\bar{d}$ de référence, nous avons bien $r(\bar{d},k) = 0$ pour tour $k = 1, \dots, K$.

Pour chaque jour d, nous pouvons considérer l'histogramme des valeurs de $r(d,k)$ pour tous les carreaux $k = 1, \dots, K$. Nous donnons une illustration de l'histogramme obtenu pour la date d du 15 août 2020 en prenant le 22 mars 2020 comme jour de référence $\bar{d}$. **La valeur de notre indicateur pour le jour d considéré est égale à l'écart-type de cet histogramme**. Les lecteurs mathématiciens reconnaîtront que cette valeur n'est autre de la distance de Wasserstein-2 entre l'histogramme du jour d (supposé comme suivant une distribution gaussienne) et l'histogramme du jour de référence $\bar{d}$ (qui est concentré à la valeur nulle).

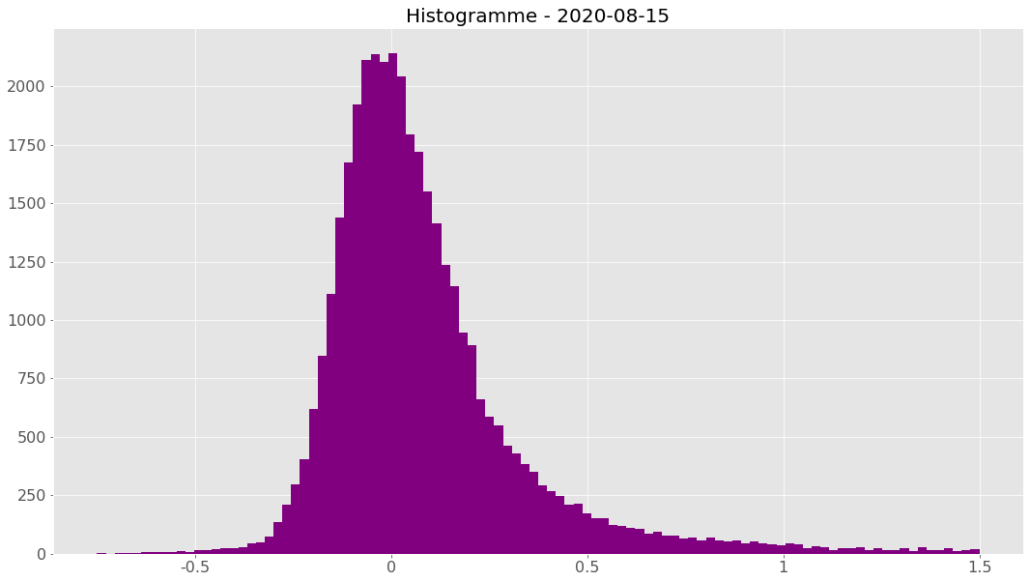


Figure A.2. Histogramme de variation relative de densité pour le 15 août 2020. L'IRD à cette date est égal à l'écart-type de l'histogramme.